Well-Behaved Variants Seldom Make the Apparatus:

Stemmata and Apparatus in Digital Research

by Barbara Bordalejo, University of Saskatchewan

To David Greetham, *in memoriam.*

> ...the editor of a traditional critical text is, in the very layout of the edition, enshrining a hierarchy of variants: those which make it onto the textual page are somehow in a different class from those which are printed in apparatus and collation.
> (D. C. Greetham, Textual Scholarship)

Introduction

This article focuses on the analysis of variants using digital resources, with a particular emphasis on the Canterbury Tales Project research and with some examples of other projects I have observed. It shows how by using a system integrating different analytical resources, specialized research tools become more accessible to scholars without the need for a background in computing.

the article on collation, co-authored with Adam Vázquez discusses what it means to collate, describes the full-text collations produced by the Canterbury Tales Project, and explains how collating in this manner presents multiple advantages over other approaches to the comparison of texts (Bordalejo and Vázquez 2020). It does not go into detail about what to do after the collation leads to variant identification. In this piece, I describe the methods the Canterbury Tales Project employs for researching the data obtained from the collation process and their uses within the framework of our project with the goal of producing the type of stemmatological research in which we focus. I describe in detail how to generate NEXUS files for use with phylogenetic analysis, the methods of phylogenetic inference, and the settings used for the project's research. I discuss how the combination of phylogenetic analysis and the specialized database, VBase, are core tools for understanding textual transmission. Finally, I conclude the article with examples of the apparatus and how they link to the stemmata, demonstrating that this research implies thousands of minute decisions, each of which carries its own weight on the final product.

The Means of Textual Criticism

There is a temptation when studying primary sources to discern and explain the most minute details of the documents at hand. It is easy to get lost within a single manuscript, allowing every mark on the page to take on a new meaning and making its representation the subject of endless revision and debate. Transcription, however, is a means of textual criticism, not its end. The preparation and process of semi-automatic computer-assisted collation are so time-consuming, particularly when working with more than fifty witnesses of significant length, that by the time a researcher emerges on the other side, she discovers herself surrounded by a sea of variants. Collation, however, is also a means of textual criticism, not its end. The analysis of variants, with all its intricacies, appears, initially, like the area in which a textual critic might focus her efforts to unveil the mysteries of the transmission of the text. The analysis of variants is one of the many

requirements in the production of editions and, occasionally, it might be the sole focus and end of textual-critical research. For many scholars, however, the analysis of variants is a means of textual criticism, not its ends. For others, notably, Wendy Phillips-Rodriguez, the production of a critical edition is not the ends of textual criticism. Phillips-Rodriguez noted this when she stated that she did not advocate the production of a new critical edition of the *Mahābhārata* but rather the exploration of the textual tradition (Phillips-Rodriguez 2007, 174).

Some might think that a research focus on philology and textual criticism is akin to falling into a rabbit hole. In fact, textual scholarship is more a lair than a hole, with interconnected galleries, tunnels leading in different directions and labyrinthian passages revealing only a few final destinations. However, with the right tools, one might be able to trace these paths, map their layout, and reach a better understanding of the lair.

Textual Communities

The current phase of the Canterbury Tales Project is supported by Textual Communities, a web-based system optimized for all aspects of the production of editions or, as stated in its launch-document, "an environment for the collaborative online creation of scholarly editions" (Robinson 2019b). Textual Communities can be used to create editions of all types of texts, but it is particularly adept at the production of editions of texts preserved in large traditions. The integration of CollateX and the Collation Editor (Smith 2019), developed by Catherine Smith (University of Birmingham), sets collation as a fundamental focus of the developers of this system.

Within Textual Communities, one can transcribe TEI-XML documents and validate them against a default DTD. The system creates two separate trees for each file, one representing the structure of the document and the other understanding the text as a communicative act (Robinson 2018; 2019a). This double tree structure allows us to retrieve the first line of the first folio in the Hengwrt manuscript or the first line of Chaucer's *Canterbury Tales*.

As explained in a separate article (Bordalejo and Vázquez 2020), the Collation Editor (Smith 2019) allows textual scholars to fine-tune the processes of regularization and alignment to produce precise collations that become the bases of other analyses and of the apparatus for the project's editions.

Extracting Data from Textual Communities

Using Universal Resource Identifiers (URIs), scholars are able to retrieve the XML apparatus from the Textual Communities databases. Some instructions on how to do this can be found as part of the Textual Communities wiki, where there are a few examples of the naming structure for the URIs (Robinson 2020).

Peter Robinson completed a new regularization of "The Wife of Bath's Prologue" in 2019, and this is currently stored in the Textual Communities database. By using a positive collation URI, scholars can extract an XML apparatus, with an optimized alignment intended for use with phylogenetic software, as seen in figure 1.

```
    <app from="10" n="CTP2:entity=WBP:line=2" to="12" type="main">
        <lem wit="Base">is right</lem>
        <rdg n="a" varSeq="1" wit="Ad1 Ad3 Base Bo1 Bo2 Bw Ch Cp Cx1 Cx2 Dd Dl Ds1 En1 En3
Fi Gl Ha2 He Hg Ht La Lc Ld1 Ld2 Ln Mc Mg Mm Ne Nl Ph2 Pn Ps Ra1 Ra3 Ry2 Se Si Sl1 Sl2 Tc1
To1 Wy">is
right<wit><idno>Ad1</idno><idno>Ad3</idno><idno>Base</idno><idno>Bo1</idno><idno>Bo2</idno
><idno>Bw</idno><idno>Ch</idno><idno>Cp</idno><idno>Cx1</idno><idno>Cx2</idno><idno>Dd</
idno><idno>Dl</idno><idno>Ds1</idno><idno>En1</idno><idno>En3</idno><idno>Fi</idno><idno>
Gl</idno><idno>Ha2</idno><idno>He</idno><idno>Hg</idno><idno>Ht</idno><idno>La</idno><idno
>Lc</idno><idno>Ld1</idno><idno>Ld2</idno><idno>Ln</idno><idno>Mc</idno><idno>Mg</idno><
idno>Mm</idno><idno>Ne</idno><idno>Nl</idno><idno>Ph2</idno><idno>Pn</idno><idno>Ps</idno>
<idno>Ra1</idno><idno>Ra3</idno><idno>Ry2</idno><idno>Se</idno><idno>Si</idno><idno>Sl1</
idno><idno>Sl2</idno><idno>Tc1</idno><idno>To1</idno><idno>Wy</idno></wit></rdg>
        <rdg n="b" varSeq="2" wit="Ha4">it were<wit><idno>Ha4</idno></wit></rdg>
        <rdg n="c" varSeq="3" wit="Ra2">it is right<wit><idno>Ra2</idno></wit></rdg>
        <rdg n="d" varSeq="4" wit="El Py">were
right<wit><idno>El</idno><idno>Py</idno></wit></rdg>
        <rdg n="e" varSeq="5" wit="Hk Ph3 Pw Ry1
Tc2">right<wit><idno>Hk</idno><idno>Ph3</idno><idno>Pw</idno><idno>Ry1</idno><idno>Tc2</
idno></wit></rdg>
        <rdg n="f" type="om" varSeq="6" wit="Ii
Ma">om<wit><idno>Ii</idno><idno>Ma</idno></wit></rdg>
    </app>
    <app from="14" n="CTP2:entity=WBP:line=2" to="14" type="main">
        <lem wit="Base">ynogh</lem>
        <rdg n="a" varSeq="1" wit="Ad1 Ad3 Base Bo1 Bo2 Bw Ch Cp Cx1 Cx2 Dd Dl Ds1 El En1
En3 Fi Gl Ha2 Ha4 He Hg Hk Ht Ii La Lc Ld1 Ld2 Ln Mc Mg Mm Ne Nl Ph2 Ph3 Pn Ps Pw Py Ra1
Ra2 Ra3 Ry1 Ry2 Se Si Sl1 Sl2 Tc1 Tc2 To1
Wy">ynogh<wit><idno>Ad1</idno><idno>Ad3</idno><idno>Base</idno><idno>Bo1</idno><idno>Bo2</
idno><idno>Bw</idno><idno>Ch</idno><idno>Cp</idno><idno>Cx1</idno><idno>Cx2</idno><idno>Dd
</idno><idno>Dl</idno><idno>Ds1</idno><idno>El</idno><idno>En1</idno><idno>En3</idno><idno
>Fi</idno><idno>Gl</idno><idno>Ha2</idno><idno>Ha4</idno><idno>He</idno><idno>Hg</idno><
idno>Hk</idno><idno>Ht</idno><idno>Ii</idno><idno>La</idno><idno>Lc</idno><idno>Ld1</idno>
<idno>Ld2</idno><idno>Ln</idno><idno>Mc</idno><idno>Mg</idno><idno>Mm</idno><idno>Ne</idno
><idno>Nl</idno><idno>Ph2</idno><idno>Ph3</idno><idno>Pn</idno><idno>Ps</idno><idno>Pw</
idno><idno>Py</idno><idno>Ra1</idno><idno>Ra2</idno><idno>Ra3</idno><idno>Ry1</idno><idno>
Ry2</idno><idno>Se</idno><idno>Si</idno><idno>Sl1</idno><idno>Sl2</idno><idno>Tc1</idno><
idno>Tc2</idno><idno>To1</idno><idno>Wy</idno></wit></rdg>
        <rdg n="b" type="om" varSeq="2" wit="Ma">om<wit><idno>Ma</idno></wit></rdg>
    </app>
```

Fig. 1. “The Wife of Bath Prologue” XML apparatus, extracted from Textual Communities. Regularized and aligned by Peter Robinson.

XML Apparatus Structure

Each word (a token in CollateX language) is assigned an even number; thus, 2, 4, 6, etc., leaving the uneven numbers free in case of additions. The encoding of each place of variation is given as <app from="10" n="CTP2:entity=WBP:line=2" to="12" type="main">. In this particular case, “10” refers to what would be position five within the line. It could have said “from=”10” and “to=”10”, which would have referred to a single word in the base text. Instead, “from=”10” and “to=”12,” corresponding to positions 5 and 6 in the line. By contrast, the following word, “ynough” appears on its own. <app from="14" n="CTP2:entity=WBP:line=2" to="14" type="main">.

Between the element attributes that determine the position in the line, there is another one, n="CTP2:entity=WBP:line=2," pointing out at the exact place in which this is occurring in the Wife of Bath's Prologue. The attribute refers to line 2 of the entity "Wife of Bath's Prologue," which is part of the Canterbury Tales Project phase 2. The final attribute within the app entity is type="main." This indicates that the section in question is present as opposed to text not present (type="om") or to a physical gap in the document (type="lac"). The next line of code in figure 1, shows the <lem> element with an attribute "Base," the lemma against which all witnesses are compared and which has been tuned during the collation process to present the words in positions 10 and 12 together as a phrase. Thus, the lemma becomes "is right," rather than being split into two lemmata, "is" and "right." This alignment is an improvement both for the phylogenetic analysis and for the readability of the generated apparatus.

The following element is a reading proper, <rdg>, which has the attribute n (with values a, b, c, etc.) and varSeq (with values 1, 2, 3, etc.). Both attributes are generated by Smith's Collation Editor, the tool employed for both regularization and alignment. Immediately, this is followed by the attribute "wit" and with the sigils of the different witnesses as value. The content of the element <rdg> is, in this case, two words, the reference reading from the base text, "is right," and the sigils of the individual witnesses marked up with the element <idno>. Each <idno> is followed by the other readings in this place of variation.

Stemmatology

After obtaining a reliable XML apparatus, which has undergone both regularization and alignment for stemmatological analysis and for optimal human reading upon conversion, has been produced, this can be processed within Textual Communities to obtain a NEXUS file. The operation is carried out within Textual Communities by choosing the options Manage => Collation => Convert Collation output to NEXUS file.

NEXUS File and Variant Matrix

NEXUS files are commonly used in bioinformatics and are a standard format for phylogenetic software (Maddison, Swofford, and Maddison 1997). Each file has several sections, formally known as blocks. Every NEXUS file has, at least, a taxa block (a taxon is an organism recognized by researchers as a unit) and a data block. Our groupings refer to texts, not organisms, so each taxon is represented by a sigil corresponding to one of our witnesses. A typical NEXUS file (Bordalejo and Robinson 2020b) used by Canterbury Project researchers will have a block called statelabels, which records all the variants at every place of variation within that section of the text. Thus, for example:

```
3277 CTP2_entity_WBP_line_484_2_2 i and for_i
3278 CTP2_entity_WBP_line_484_4_12 made_hym_of_the_same
3279 CTP2_entity_WBP_line_484_14_14 wode hode clothe wede
3280 CTP2_entity_WBP_line_484_16_18 a_troce a_croce a_cote a_groce an_hode
```

Although the syntax is not designed for human reading, it is not difficult to understand. Each entry has an individual identifier in consecutive order, followed by the line indicators and the position within the line. The example above can be read as follows: these are characters 3277, 3278, 3279 and 3280 of the Wife of Bath's Prologue's line 484 locations 2 to 18 within phase 2

of the Canterbury Tales Project. The positions within line WBP484 are given in even numbers. This is part of the processing by Catherine Smith's Collation Editor (Smith 2019). Because of the limitations of CollateX, which does not have regularization or alignment facilities, Textual Communities has also embedded a version of Smith's Collation Editor. Catherine Smith, working for the Institute for Textual Scholarship and Electronic Editing at the University of Birmingham, developed the Collation Editor to regularize and align the Greek New Testament. For anyone familiar with the *Editio Critica Maior* (Aland, Strutwolf, and Universität Münster 1997), the use of only even numbers is understandable.

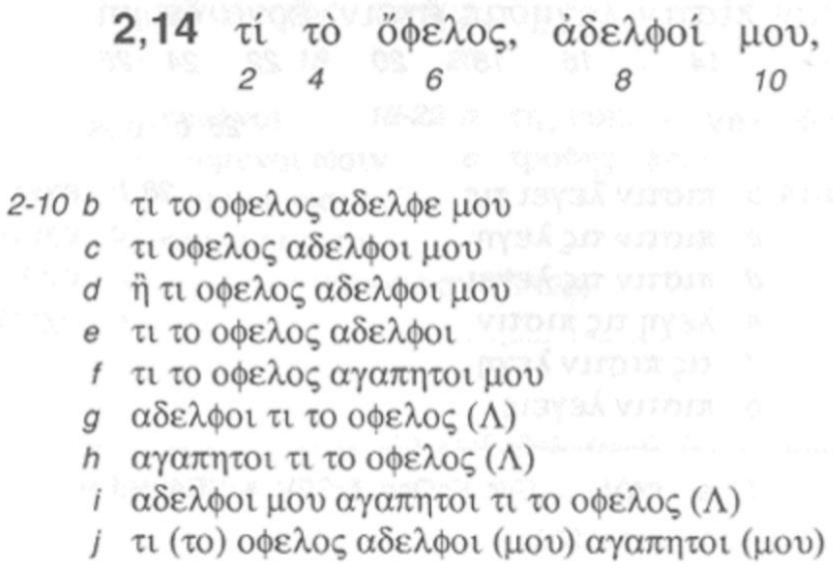

Fig. 2. The text and condensed apparatus of the *Editio Critica Maior*

The concept is that variants are given in reference to the text present in the base text using the even number system, but when text not present in the base text comes to light, this is labelled with odd numbers. Smith has followed the same architecture in her program, not only because her software is in use at the Institute for New Testament Research in Muenster for the continued work on the *Editio Critica Maior* and the Nestle-Aland edition, but also because the idea is both sound and successful in practice.

Returning to the specifics of the previous example (Figure 1), the places of variation are separated by a space in the nexus file, while phases are presented as a unit by using the underscore. This means that place of variation 2, the first in this line, has three variants:

I
And
For I

While place of variation 14, the third in the line, has four variants:
wode
hode
clothe
wede

The second place of variation in the line is a phrase, “made hym of the same,” formed of places of variation 4 to 12. This means that the words are in the second to the sixth positions in the line.

For the Wife of Bath’s Prologue, there are 5463 places of variation after regularization and alignment.

The following block of the NEXUS file is the taxlabels block, where a number is assigned to each witness, starting with 0 for the Base text and going to 88, which corresponds to Wynkyn de Worde’s edition.

The final section of a Canterbury Tales Project NEXUS is the variant matrix, just called Matrix in our file.

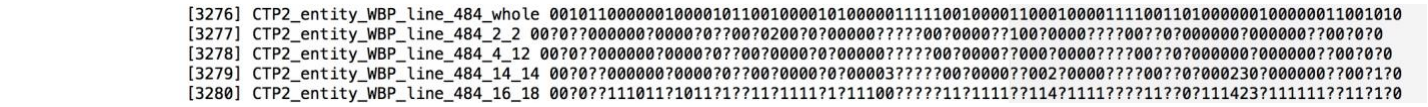

Fig. 3. Detail of the matrix in the Wife of Bath’s Prologue NEXUS file.
.
The matrix is difficult to read for humans, though not impossible. Each variant has been assigned a number. For example, in places of variation 16 and 18, which have been aligned together, there are four phrase variants:

a troce
a croce
a cote
a groce
an hode

Each of these variants gets assigned a number 0 (a troce), 1 (a croce), 2 (a cote), 3 (a grote), 4 (an hode). Text not present is marked with a question mark. Consider the following:

> [3280] CTP2_entity_WBP_line_484_16_18
> 00?0??111011?1011?1??11?1111?1?11100?????11?1111??114?1111?
> ???11??0?111423?111111??11?1?0

Witnesses that agree with the base text are represented as 0. By counting the position and reconciling that with the position of each witness sigil in the taxlabel block, one could figure out which witnesses agree with the base text, which with reading 1, 2, 3 or 4. Although not impossible to interpret, the task is time-consuming and not advisable. For human reading, it is much easier to present something more like a traditional print apparatus (whether positive or negative). The NEXUS file is needed to process it with Phylogenetic Analysis Using Parsimony and Other Methods (Chaucer 2020), henceforth PAUP, or some other bioinformatics software. The NEXUS file for “The Wife of Bath Prologue” used in these examples was generated from the apparatus produced after Robinson’s regularization (Bordalejo and Robinson 2020a).

Phylogenetic Analysis Methods

For textual scholars handling relatively large amounts of data, the use of computers might be quite obvious. And yet, only a small number of textual scholars make regular use of any kind of computer-assisted stemmatological method. Howe et al. outline the reasons to employ these as part of one's research:

> These methodologies collectively offer several advantages to textual scholars. The use of computers means that the dataset can be sensitively yet comprehensively handled, which is particularly important if it comprises copies of a long text, for which a manual analysis may prove unwieldy. Multiple approaches can be applied to the same dataset for the purposes of comparison and testing. Perhaps most importantly for scholars of vernacular traditions, not all phylogenetic analyses are tied to the assumption that a single ancestor is responsible for each extant copy and some copies are capable in principle of showing multiple affiliations. (Howe, Connolly, and Windram 2012, 56)

The ability to comprehensive handle data is one of the most significant aspects of the use of this software, because of the computer's capacity to accurately record the enormous amounts of data derived from a single textual tradition (for a comparison between manual and computer-assisted methods, see Bordalejo and Vázquez 2020). And yet, the open possibility of continued testing of different approaches (including different models of phylogenetic inference) cannot be underestimated. Through the years, the Canterbury Tales Project researchers have done precisely this by employing multiple methods with the same dataset. Peter Robinson experimented with other software, such as SplitsTree (Huson and Bryant 2006), which I also tested for my work on the order of the *Tales* (Bordalejo 2003). PAUP (Swofford 2003) produced clearer results than SplitsTree and has become the standard tool for Canterbury Tales Project research. Although serious effort has gone into the development of software specializing in literary and historical textual transmission, neither RHM (Roos and Heikkila 2009) nor SemStem (Roos and Zou 2011) has yet been made widely available Tools to facilitate computer-assisted stemmatological analysis have been incorporated into Textual Communities, while Robinson continues conversations for the integration of stemmatological software into the system.
The software aims to present a seamless interface to the users with the file conversion happening in the background in a similar way to which it currently converts the apparatus output into a NEXUS file. The plan is to integrate the software within Textual Communities to facilitate its use through a single interface and to produce results ready for display as part of digital editions produced using the Textual Communities API.

To detail the debate over the validity of stemmatological methods in textual criticism would be beyond the purview of this paper, however, a brief account of the ongoing discussion around such methods demonstrates the impact/importance of our work. Although stemmatological methods have been used successfully to explore diverse textual traditions (Barbrook et al. 1998; Spencer, Bordalejo, Robinson, et al. 2003; P. Robinson 2003; Chaucer 2004; 2006; Eagleton and Spencer 2006; H. F. Windram et al. 2008; Phillips-Rodriguez 2010; P. M. W. Robinson 2012), the singular idea that the relationship between phylogenetic analysis and manuscript transmission was of analogy, rather of identity, prompted targeted essays focusing on the relationship between

the two (Mooney et al. 2004). Independently of its justification, the use of tools originally developed for a specialized field and later implemented in another has not been without criticism (Robins 2007, Hanna 2000, Cartlidge 2001, Alexanderson 2018). For this reason, various experiments were carried out with artificial textual traditions (Spencer et al. 2004; Roos and Heikkila 2009) and, when these were still not considered enough, a response to the criticisms was published (Howe, Connolly, and Windram 2012) followed by further theoretical explanations (Bordalejo 2016). Those interested in the use of phylogenetic analysis or other stemmatological tools would acquire reasonable knowledge from the texts mentioned above.

The success of phylogenetic methods with various textual traditions has been paralleled in other fields, most notably anthropology (Tehrani and Collard 2002; Tehrani and Collard 2009), folklore (Tehrani 2013; Tehrani, Nguyen, and Roos 2016), and archeology (Mendoza Straffon 2016). These applications, beyond molecular evolution, with its expansion to non-biological fields, gave rise to the concept of "phylomemetics" (Howe and Windram 2011).

I have written about the theory and practice of using phylogenetic analysis in the research context of medieval manuscript traditions (Bordalejo 2003, 90-112). However, since the work remains unpublished, it seems pertinent to summarize some important matters in this piece. Phylogenetic software looks into the nucleotide sequences to isolate the three-letter words that encode individual amino acids and how the copying process sometimes results in the loss of a nucleotide that gets replaced by a different one giving rise to a mutation. A mutation, if it were to give rise to a successful advantage, would be inherited and become a feature (Bordalejo 2003, 91-92). The software can express its results as networks or trees. Stemmatologists might tend to favour tree-building software because its output appears closer to that of conventionally constructed stemmata. The type of data used and the tree-building method sort phylogenetic software into categories.

Data Handling: Distance vs. Discrete

The data for use with phylogenetic software can be approached directly by structuring the data as a NEXUS file, as described above (this is what the Canterbury Tales Project does with textual data); or data can be converted into a distance matrix, as was done with the *Canterbury Tales* tale-order data (see below). When the only step for processing is the restructuring of data, one talks about a discrete method. When the data is processed and converted into a distance matrix, one talks about distance methods.

Page and Holmes explain that "…[d]istance methods are based on the idea that if we knew the actual evolutionary distance between all members of a set of sequences, then we could easily reconstruct the evolutionary history of those sequences" (Page and Holmes 1998, 179). Unweighted pair-group method using arithmetic averages (UPGMA), Least Squares (LS), Minimum Evolution (ME), and Neighbor Joining (NJ) are all distance methods (Nei and Kumar 2000, 87-113). For each pair of taxa (or witnesses), the evolutionary distance, which is a measure of genetic diversity, is calculated. The constructed tree considers the relationships between the distance values (Nei and Kumar 2000, 87).

Although it might seem obvious that by removing the conversion into a distance matrix, one might also remove another layer for the possible introduction of errors or the mediation of models that can impact the data, not all data is liable to simple restructuring. Such was the case of the tale-order data I was analyzing as part of my NYU doctoral thesis (Bordalejo 2003). The tale-order work was based on my recoding of the charts by John Manly and Edith Rickert (Manly and Rickert 1940), but these data required conversion prior to analysis. Matthew Spencer, who was also part of the STEMMA Project, suggested the use distance methods and was the first to convert my tables so phylogenetic software could be tested with non-textual data. This serves as an example of the use of distance methods when direct data restructuring is not possible.

Spencer used breakpoint distance method (BP) which worked because the witnesses shared a significant number of missing items and also because there were fewer common items missing between any given pair. Spencer's algorithm generated upper and lower bound data which differ from each other in that "[t]he lower limit occurs when no common items were lost, and the upper limit is approached if there are many lost common items" (Spencer et al., unpublished). However, breakpoint distance "is only reliable when the number of transpositions is small" (Spencer, Bordalejo, Wang, et al. 2003, 102). At the time, Wang and Warnow had just devised Inverse of Expected BreakPoint Distance (INBP), which seemed better suited for the tale-order research (Wang 2001). Both methods are described in our article, "Analyzing the Order of Items in Manuscripts of *The Canterbury Tales*" (Spencer, Bordalejo, Wang, et al. 2003), where ME trees based on this data are presented. Fuller results of the tale-order analysis are presented in my NYU doctoral thesis which concludes that there is an undeniable coherence between tale-order and textual transmission in the tradition which suggests that the order was more often than not copied from an exemplar while, in few occasions, it was purposely altered by scribes or their supervisors (Bordalejo 2003. 190 and ff.). Because of the nature of the data, trees were built using ME and NJ, both phylogenetic inference methods that accept distance values as data input.

Discrete methods, by contrast, use structured data without the extra step in processing. Thus, they are one step closer to the data than distance methods. Some data, like the tale-order data, because of their nature, must be encoded before processing. The NEXUS file based on the Wife of Bath's apparatus is a restructuring of the data to be processed by phylogenetic software, but the data is not changed by such restructuring. Discrete methods "endeavour to avoid the loss of information that occurs when sequences are converted into distances" (Page and Holmes 1998, 187), which means that another degree of separation between the data and the resulting tree is avoided. Maximum Parsimony (MP) and Maximum Likelihood (ML) are discrete methods of phylogenetic inference. These methods differ on how they choose the trees they present:

> The two major discrete methods are maximum parsimony and maximum likelihood. Maximum parsimony chooses the tree (or trees) that require the fewest evolutionary changes. Maximum likelihood chooses the tree (or trees) that of all trees is the one that is most likely to have produced the observed data (Page and Holmes 1998, 187).

In my previous study, I explain that although these methods are particularly well-suited for dealing with textual variation, the fact that MP searches for the trees with the least number of

changes is liable to present a simplified version of what might be a more complex tradition (Bordalejo 2003, 93). However, there is a more significant problem with MP. Nei and Kumar synthesize it as follows:

> If there are no backwards and no parallel substitutions (no homoplasy) at each nucleotide site and the number of nucleotides examined (*n*) is very large, MP methods are expected to produce the correct (realized) tree (Nei and Kumar 2000).

Homoplasy refers to both parallel and convergent evolution, both of which are cases of independent development of the same features. Textual scholars are familiar with this phenomenon during which different scribes in completely separate occasions introduce the same change to the text. Manly and Rickert call it agreement by coincidence. This type of inference does not work well with highly contaminated traditions. Fortunately, for 15th-century witnesses of the *Canterbury Tales* contamination a less significant issue that it is in larger traditions with a life-span of several centuries like that of the *Mahābhārata* or the Greek New Testament. Models of phylogenetic inference that could cope with contamination and coincidental agreement would be advantageous for both large classical and medieval textual traditions.

The Canterbury Tales Project

In 1999, *The General Prologue on CD-ROM* (Solopova 2000) presented trees that were produced using SplitTrees and that were informative about an early split in the textual tradition from which Robinson hypothesized the alpha hyparchetype (a lost manuscript from which roughly half of the tradition descended).

Despite this breakthrough, the control offered by PAUP was unparalleled, and its results were consistent with aspects of the textual tradition confirmed independently. Take, for example, the fundamental groups proposed by Manly and Rickert:

Group *a*: Cn Dd En1 Ds Ma
Group *b*: He Ne Cx1 Tc2
Group *c*: Cp La Sl2
Group *d*: En2 Ll1 Lc Mg Pw Mm Ph3 Ry2 Ld2 Dl Ha2 Sl1
Independent pairs: Ad3 and Ha5, Bo1 and Ph2, En3 and Ad1, Mc and Ra1, Ps and Ha1, and Ra2 and Ht

These groupings and most of the pairs are confirmed by analysis carried out by members of the Canterbury Tales Project (Chaucer 1996; Robinson 1997; Robinson 2003; Bordalejo 2003; Chaucer 2004). Phylogenetic software confirms part of Manly and Rickert's manual analysis of the *Tales*. The phylogenetic trees offer enough new avenues of enquiry to open paths for further research. There are two main conclusions from our analyses that improve or correct Manly and Rickert:

1) The tales and sections did not circulate independently; witnesses share both for text and non-textual elements the same overall relationships.

2) Hg, El, Ch, (Ad3, Ha5), (Bo1, Ph2), (En3 Ad1), Mc and Ra1, Ps and Ha1, and Bo2 and Ht represent independent lines of descent from the archetype. We call them the o witnesses (Robinson 1997, 80).

Neither of these two corrections to Manly and Rickert signal incompetency nor carelessness. They were both excellent editors and researchers. The hypothesis of the independent circulation of the *Tales* appears to have support from the fact that the work was left unfinished and some units shifted positions, but both the textual and non-textual data indicate that the *Canterbury Tales* circulated as a book rather than in booklets (Chaucer 1996; Bordalejo 2002; 2003; Chaucer 2004; 2006).

The Canterbury Tales Project's editions present Variant Maps, representations of the relations among the witnesses produced using phylogenetic software and displayed as unrooted tree-like graphs. Some scholars might consider this controversial, but I want to state it here as clearly as possible: the unrooted graphs, which we call Variant Maps in our editions, are stemmata. They are based on data informed by editorial judgement at every point and show genetic relations among textual witnesses. Although the stemmata could be rooted, a root is unnecessary because it would not change the relationships between the nodes (for more information on the fact that changing the root of a tree does not change the relationships between the witnesses see Robinson and O'Hara, 1993). Robinson and I deliberately choose not to create a visualization reminiscent of manual stemmata. It is not necessary as Robinson and O'Hara showed almost thirty years ago.

Why Maximum Parsimony

Because MP does not require further data processing, it seems preferable to other approaches for use with textual variation. The underlying model of phylogenetic inference, seeking the most parsimonious tree, "...creates a tree that represents the smallest overall number of independent mutations..." (Howe, Connolly, and Windram 2012, 63). Elsewhere, I explain that these trees should not be expected to conform to the historical reality of a manuscript tradition, as they can only take into account the input data, which is generally textual rather than extra-textual (Bordalejo 2016, 568).

As methods were tested, parsimony became our choice because ML required both more time and computer resources while not offering significantly better results. For a tradition with fewer witnesses, one can use MP, and PAUP will do an exhaustive search for all the possible trees before settling into the most parsimonious one. However, above a certain number of witnesses, it is better to start with a heuristic search:

> A provisional MP tree is first constructed by using a procedure called step-wise addition algorithm, and this provisional MP tree is then subjected to some kind of branch swapping to find a more parsimonious tree (Nei and Kumar 2000).

The Canterbury Tales Project routinely uses heuristic searches for the production of stemmata. These searches are complemented by comparing them with consensus trees when more than one equally parsimonious tree is found. When a section of a tree appears surprising, bootstrapping, a sampling technique that is repeated one hundred times to offer a percentage result in which

higher numbers point towards higher reliability (for more details see Higgs and Attwood 2005, 169 and ff.), is used to clarify the witnesses' support.

The caveat for the use of MP is the problem of agreement by coincidence that, in large enough numbers, would produce an incorrect topology for the tree. One has to look elsewhere for a possible solution. T-REX is a webserver for the inference, validation, and visualization of phylogenetic trees developed by members of the Department of Computer Sciences at the Université du Québec à Montréal and which deals with the issue of lateral gene transfer (Boc, Philippe, and Makarenkov 2010; Boc, Diallo, and Makarenkov 2012). Testing T-REX with textual traditions opens the possibility of solving a significant problem within computer-assisted stemmatology. This would have at least as much impact as the application of Chi-Square for the detection of a change of exemplar (Windram, Howe, and Spencer 2005; Phillips-Rodriguez, Howe, and Windram 2009).

Using phylogenetics to explore the tradition

After the apparatus data is formatted into a NEXUS file, it can be uploaded to PAUP (or another phylogenetic application accepting the same type of format). The principles outlined above are implemented, setting the software to seek MP trees using heuristic searches.

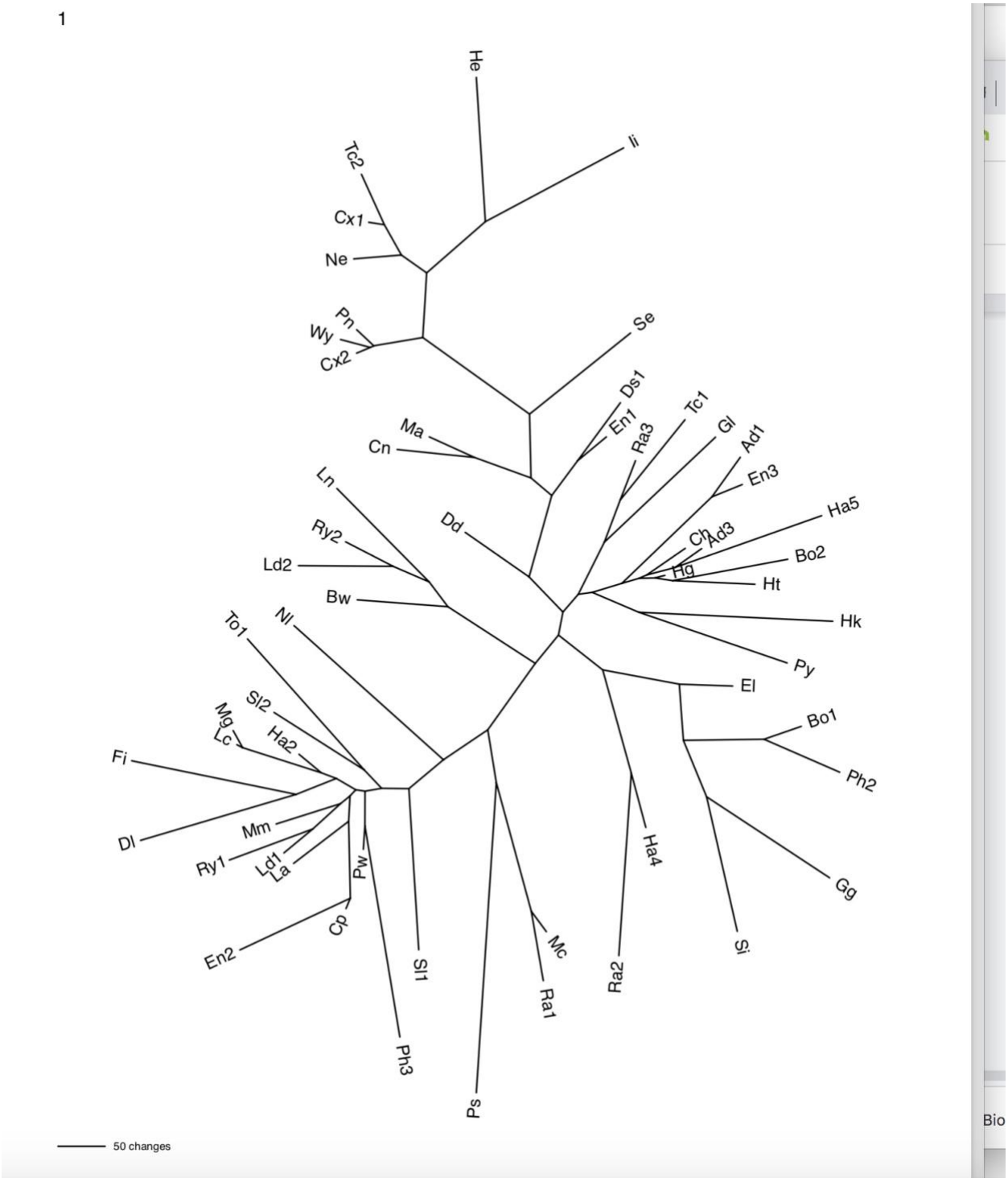

Fig. 4. Stemma of lines 1 to 400 of “The Wife of Bath’s Prologue.”

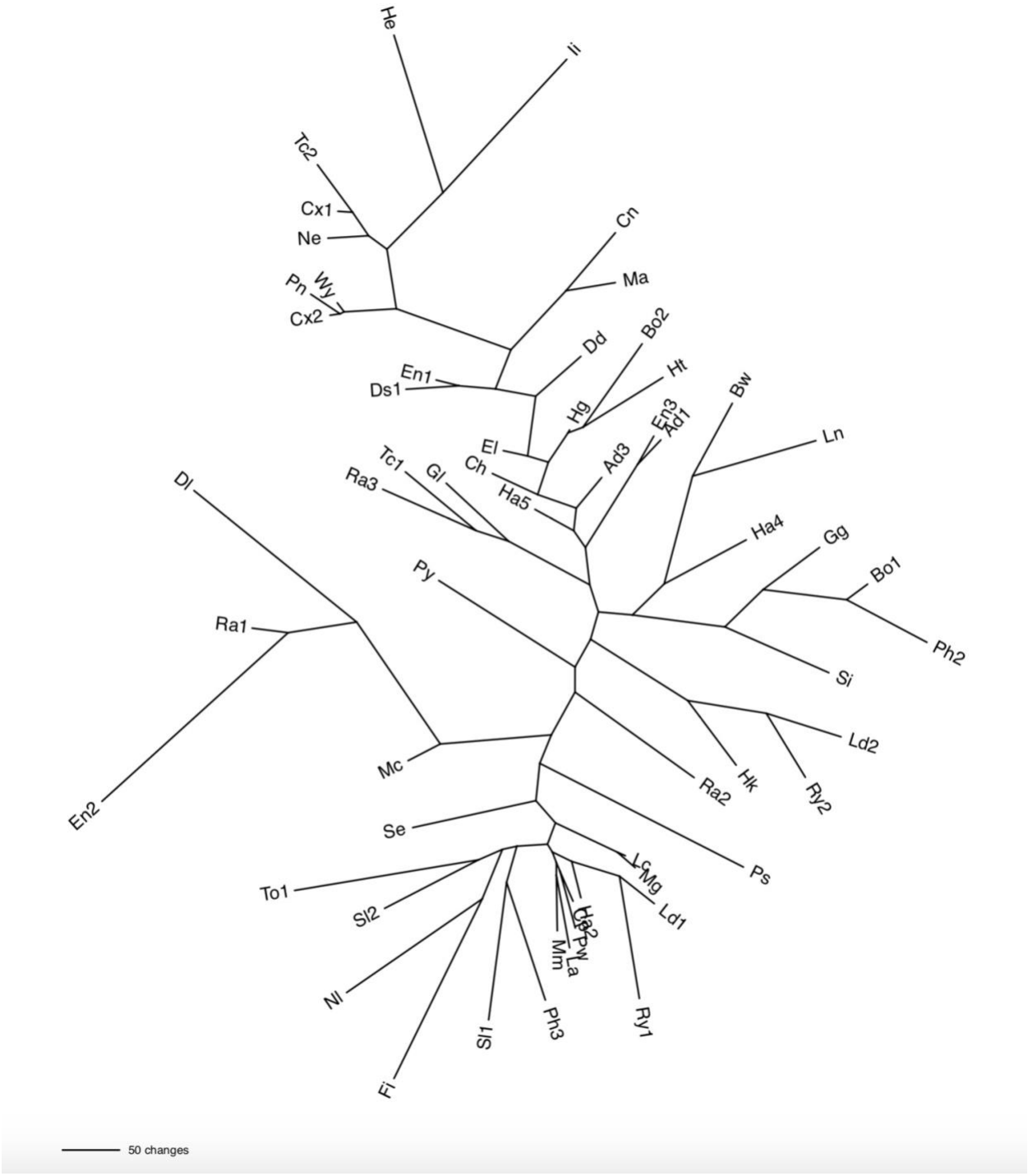

Fig. 5. Stemma of lines 401 to the end of "The Wife of Bath's Prologue."

For "The Wife of Bath's Prologue," where are two separate stemmata because the project's research indicates that the Ellesmere manuscript (El) changes exemplars around line 400. This became clear during the research that led to the publication of *The Wife of Bath's Prologue on CD-ROM* (Chaucer 1996; Robinson 1997, 79), and was independently confirmed by the chi-square analysis carried out by Windram, Howe, and Spencer, who point out the exact place of

maximum chi-square value is character 3384, "which corresponds to line 404 in the text and is the location most likely to be the site of manuscript recombination" (Windram, Howe, and Spencer 2005, 194–95). This example shows very clearly the change of Ellesmere's habitual position within the stemmata (clustering close to Hengwrt [Hg] and Christ Church [Ch]) as it does in figure 5, to branching with Bo1, Ph2, Gg and Si (figure 4). For most scholars, just the shift from one place to another is remarkable. However, Ellesmere grouping with those manuscripts is the same pattern of variation found in "The Squire's Tale" (Bordalejo 2002, 200–3). Gg also has the tendency to shift positions in the stemmata, suggesting either a contaminated exemplar or multiple sources for the manuscript resulting in conflation. Both Gg and El share a relatively high number of this characteristic agreement below the archetype, an agreement that is both genetic and conflicts with the other textual characteristics these manuscripts present.

Manly and Rickert, for all their laboriously accurate work on the *Canterbury Tales* (Manly and Rickert 1940) were unable to understand major sections of the tradition, particularly archetypal variation and the internal relationships within the *d* group. The data is too vast for humans to classify, understand, and draw sound conclusions from. Computational methods can sort and classify accurately and allow scholars to concentrate on interpretation. Had Manly and Rickert had an automatic system to sort variants they would have been unlikely to have been confused by the witnesses carrying archetypal variation (Bordalejo and Robinson 2019). However, despite their correct identification of four manuscript groups (*a*, *b*, *c*, *d)* based on their manual analysis, Manly and Rickert were not able to make sense of archetypal variation retained in separate lines of descent, a series of witnesses which Robison has termed O, and which do not represent a genetic group.

VBase

No matter how clear our understanding of how phylogenetic software works or what inference model underlies our results, it would be foolish to accept the resulting trees with blind confidence. For this reason, our editions include VBase, a variant database that allows us to perform complex variant distribution queries. These queries help us further analyze our stemmata and explain why the phylogenetic software rendered a given tree in a particular way.

The Variant Map for line 65 of "The Miller's Tale" shows an odd distribution in which Ellesmere agrees with the *b* group (Cx1, Ne, He, and Tc2) against both Hengwrt and Christ Church. Our experience is that these three manuscripts (El, Hg, and Ch) form a compact trio that often appears with other O witnesses and, since silk/grene are not readings that arise simply by mistake, further exploration is required (see figure 6). VBase can retrieve answers to highly sophisticated questions, which might contribute to the assessment of whether there is more to this place of variation.

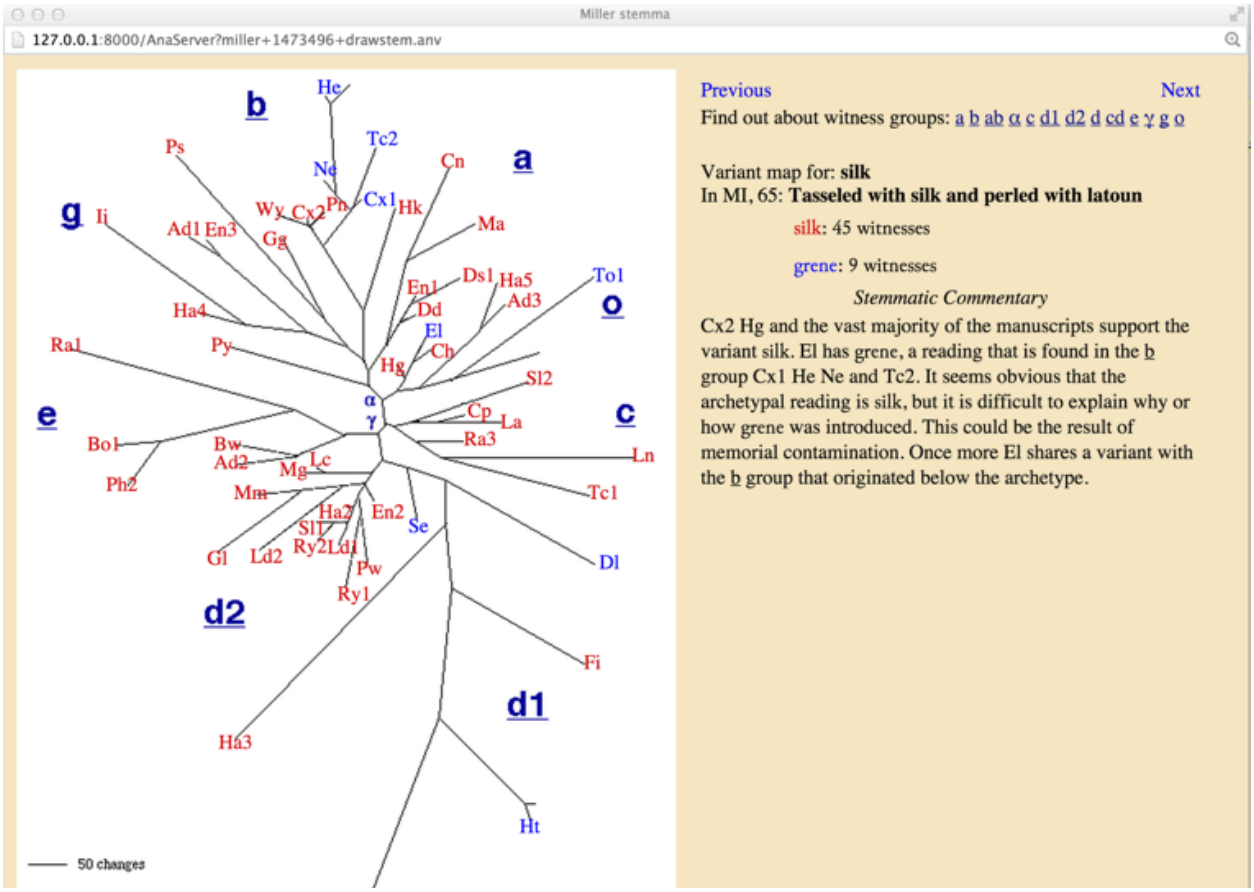

Fig. 6. “The Miller’s Tale” Variant Map (Chaucer 2003).

One might want to start with a relatively simple query that retrieves all the variants that support *b* as a genetic group. In order to establish which variants determine the *b* group, one must carry out a search, as illustrated in figure 7. In our edition, some searches are preset (and additional information on them is offered) to facilitate the use of VBase for anyone without an in-depth knowledge of the textual tradition of the *Canterbury Tales*.

There are three main steps for the retrieval of the *b* variants:

1. Eliminate archetypal readings, expressed in <14 of all and in two or more of Hengwrt, Ellesmere and Christ Church. That is, they should appear in fewer than two of the three or <2 of Hg El Ch (these witnesses share archetypal variation, and the *b* variants arose below the archetype).
2. Distinguish from the *a* group by excluding variants present in more than two of one of the *a* subgroups, expressed as <2 of En1 Ds1 Dd; and any variants in more than one of the other *a* subgroup (<2 of Ma Cn Dd). Notice that one witness, Cambridge CUL Dd 4.24, appears twice because it shares variants with both subgroups of *a* (*b* must have variants exclusive to itself to be considered a genetic group and these cannot be shared with another genetic group).
3. The separation of *b* is established by those variants that follow the above conditions while appearing in three *b* witnesses or in >2 of Cx1 He Tc2 Ne Ox1.

Fig. 7. VBase preset search for *b* group variants (Chaucer 2003).

The evidence indicates that *b* is a genetic group: there are 222 variants shared by the b witnesses not present in a significant number of the rest of the witnesses. VBase makes it possible to conduct all sorts of specialized queries. If one were curious as to the number of variants that Ellesmere shares with *b* against Hengwrt and the rest of the textual tradition, one could simply adjust the query, as seen in figure 8. VBase returns 32 places of variation showing that Ellesmere, one of the most important manuscripts of the *Tales* and a manuscript that is generally in agreement with Hengwrt and Christ Church, preserves 32 instances of variation likely to have originated below the archetype and linking it to one of the most textually removed from the origin of the tradition. I will not argue here about the reasons for this, discussed elsewhere (Chaucer 2004, Stemmatic Commentary, MI65). This is merely an example of how VBase can be used to explore questions related to witness relationships.

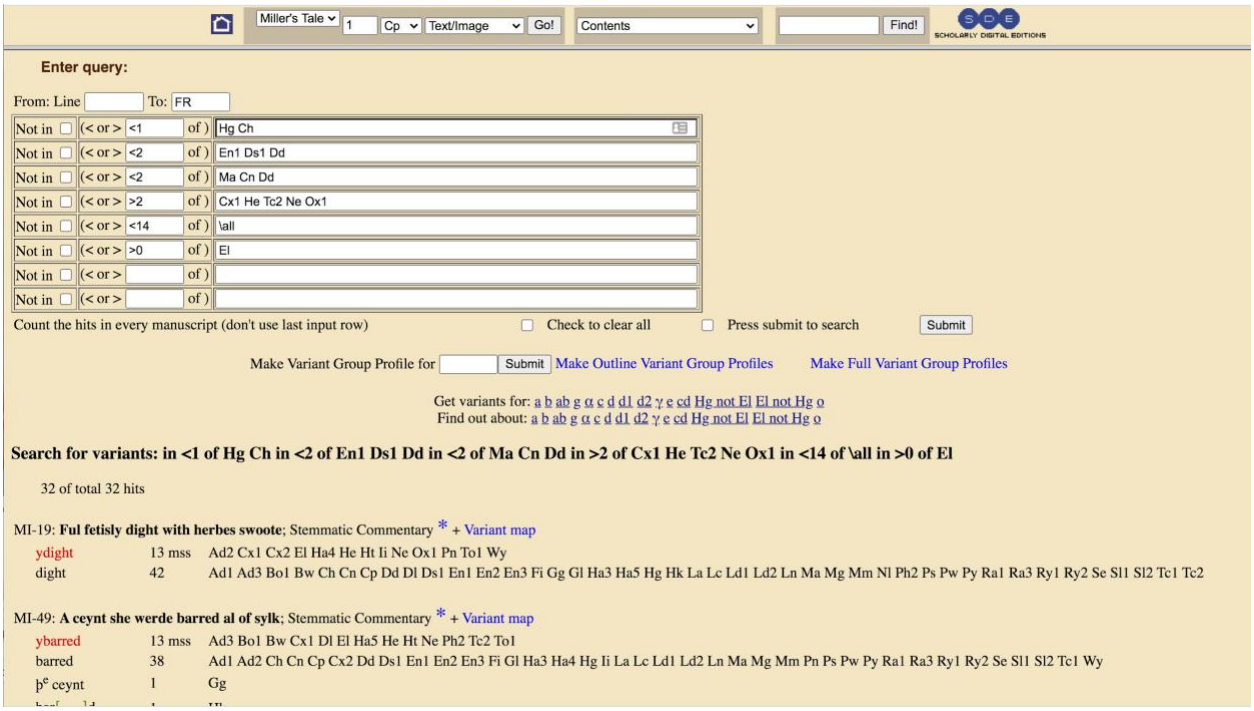

Fig. 8. VBase modified search to isolate variants shared by Ellesmere and the *b* group (Chaucer 2003).

VBase is instrumental in helping us understand the variation on which PAUP has based every section of the tree. Manly and Rickert's correct assessment of the fundamental witness groups shows that it is possible to observe and analyze data and make the correct deductions without the help of a computer. But further investigation, such as in my query trying to isolate the variants in which Ellesmere agrees with *b*, requires more precise tools.

Using a combination of PAUP and VBase, we have been able to understand the O witnesses: not a genetic group, but witnesses representing independent lines of descent from the archetype. The O witnesses often preserve archetypal readings which are lost elsewhere in the tradition. These variants puzzled Manly and Rickert who were not able to correctly classify them. Again, this is not a criticism of their work, but rather evidence of the difficulties editors face when dealing with very large datasets.

Fig. 9. VBase preset search for O variants (Chaucer 2003).

To search for O variants, one sets up a query in which the target is present in at least two of Hengwrt, Ellesmere and Christ Church and in fewer than 15 of all other witnesses. Technically, all archetypal variants preserved by any witness in the tradition should be O variants but, for our purposes, O variants are those that have been preserved (often on account of their difficulty) by a few scribes in a few witnesses derived in a more or less direct way from the archetype of the tradition (Chaucer 2004). These variants, because of their distribution within many lines of descent, and their nature, are likely to be Chaucerian. There are 37 such variants in the Miller's Tale, as shown in figure 9.

VBase allows researchers to carry out complex searches asking precise questions. Some of those questions might come from hypotheses put forward by other scholars or by one's own observations of particular witnesses in the tradition. However, when PAUP presents unexpected groupings or places witnesses in surprising positions in the tree topology, VBase can help us understand what part of the data supports the visual representation and why.

The Edition Apparatus

In most cases, readers do not seek to get profoundly involved with research on textual variation, or they require a synthetic view of a particular place of variation. Our apparatus, built from our curated collations, offers various ways to approach the *Canterbury Tales* variants. Robinson and I strive to present apparatus that are readable, but we also want them to be useful beyond the purposes of our own research.

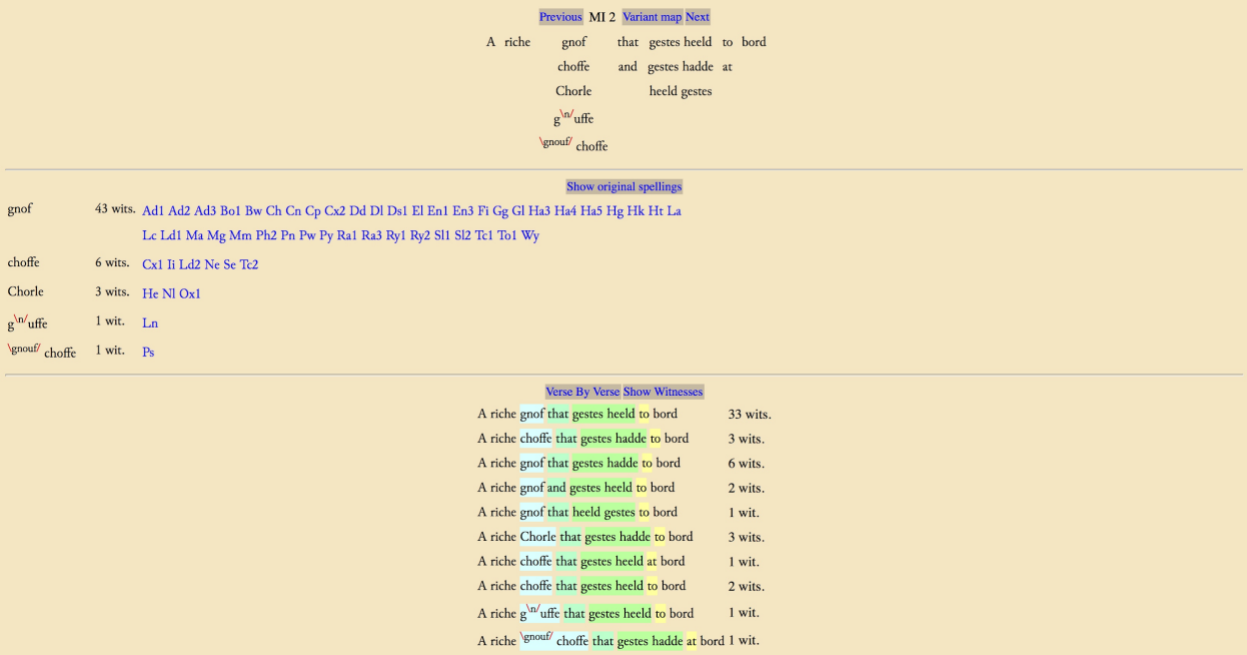

Fig. 10. The synoptic apparatus, the regularized apparatus, and the lineated apparatus. (Chaucer 2003).

Our previous editions have presented a synthetic view of the line, the synoptic apparatus (top section of figure 10). The variants appear in the middle section, which corresponds to the regularized apparatus. The last section presents an aligned lineated apparatus in which the colours suggest how it should be read vertically.

The synoptic apparatus shows all the possible variants in each place of variation but gives no indication as to which of the horizontal reading combinations is an actual line present in one of the witnesses (the lineated apparatus offers that). Instead, it presents, at a glance, the complete variation within a line.

The middle section of the apparatus shows the regularized forms of each word, although the original spellings can be shown by clicking on the link that turns them on. By having the information in three different formats, it becomes more readily comprehensible and digestible.

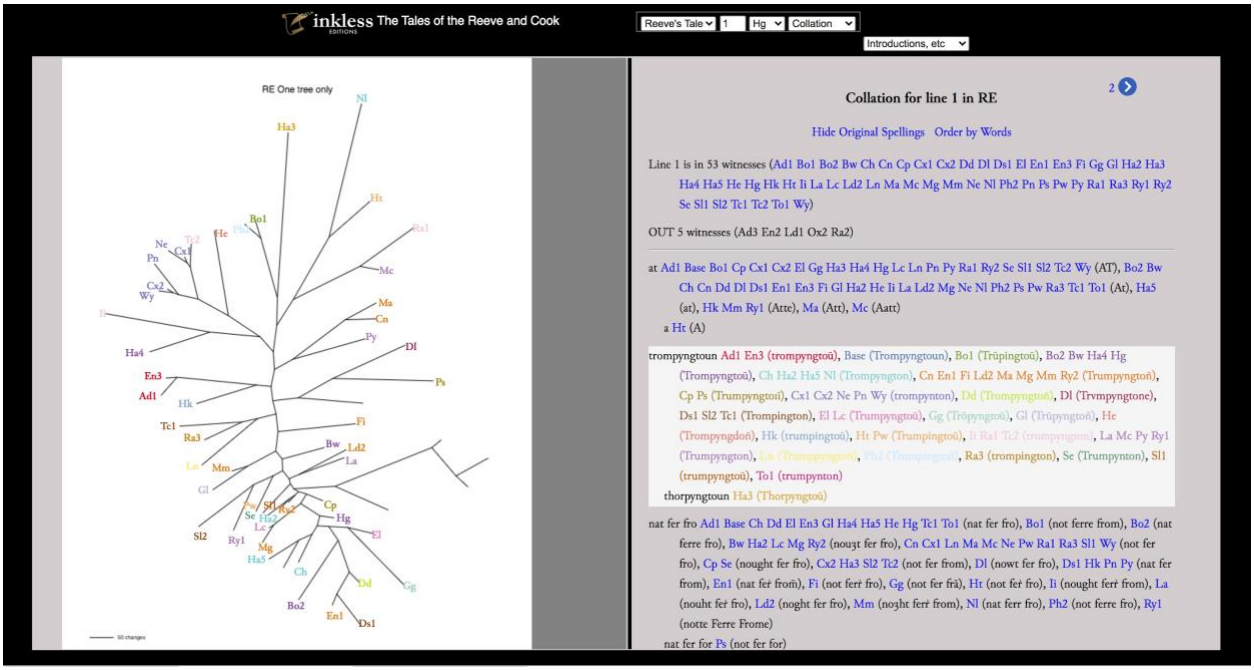

Fig. 11. The Variant Map, the regularized apparatus, and the original spellings. (Chaucer 2020).

In our latest editions, produced via the Textual Communities API, each variant is linked to the colour coded stemma. Figure 11 presents the unregularized spellings of *Trompyngtoun*, highlighting the enormous variation of spellings in toponyms, which the regularized collation, in contrast, shows to be quite consistent.

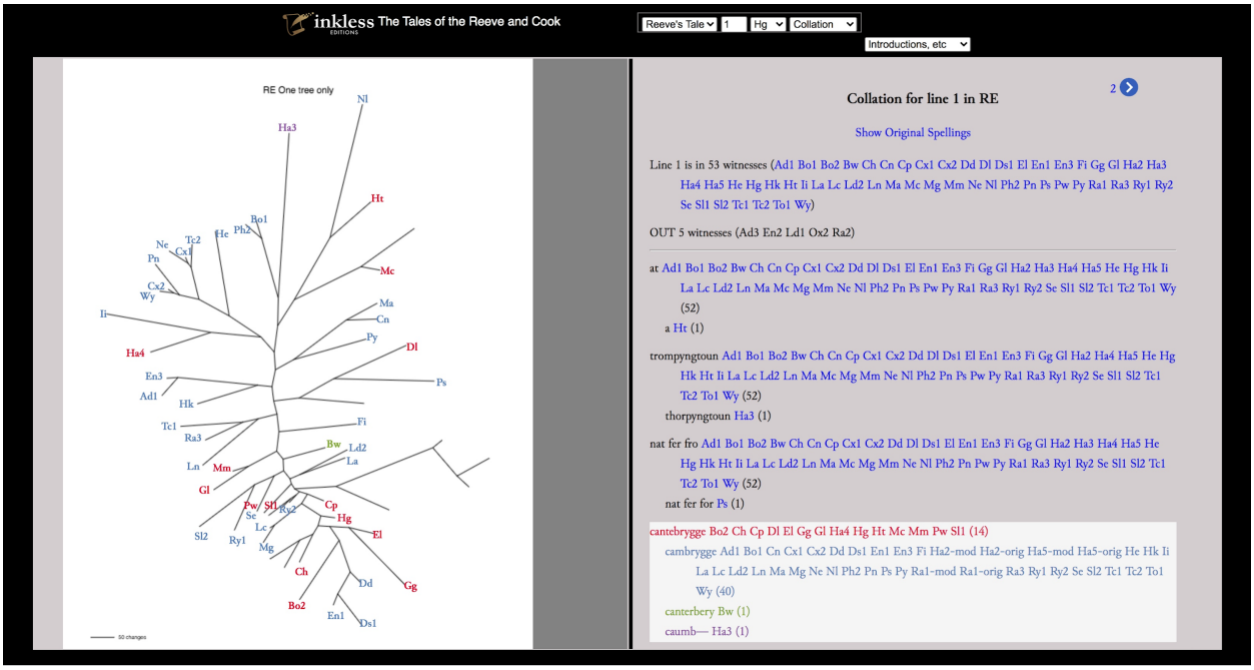

Fig. 12. The Variant Map and the regularized apparatus (Chaucer 2020).

The regularized collation of Cantebrygge/ Cambrygge, however, shows a typical example of an O variant, in which the distribution of the *lectio difficilior*, Cantebrygge appears in independent lines of descent in different sectors of the stemma.

The Ends of Textual Criticism: Understanding Textual Relations

The Canterbury Tales Project, throughout its almost thirty years of history, has pioneered approaches to digital editing, produced cutting edge research, and tested various publication platforms (for the history of the project see Robinson 2003, 2013). Peter Robinson developed collation software and three separate publication systems to fulfill the requirements of working with large textual traditions, while members of the project and other collaborators contributed to the user interfaces and designed visualizations for various aspects of the display of our editions. Today, the development of Textual Communities, implementing CollateX and the Collation Editor as well as file conversion facilities, presents the integration of robust systems while still seeking to innovate.

A significant proportion of this research could be carried out by hand, but then one would be left with little time to think through its results and attempt new approaches. The time saved by using digital tools can serve to prioritize the investigation of new ways of exploring textual traditions. Additionally, others benefit from the project's work which they can employ in their own research. The methods employed by the Canterbury Tales Project work, but they take time to learn and effort to understand. I have tried to give a detailed account not only of why the project takes these approaches but of how the methods work, so scholars interested in testing these approaches might be encouraged to try them.

A future avenue of exploration would be on the application of software that solves some of the problems with MP, namely agreement by coincidence; T-REX reputedly does this, but it remains to be tested with textual data. Given my experience with other bioinformatics software, I feel confident that this exploration, if successful, might push our understanding of contaminated textual traditions. However, it is essential to evaluate the quality of the results, particularly whether said results are good enough to warrant the effort of presenting the data in a different format, learning how to use the new software, and fully understanding its underlying model.

All of this work, from the decisions as to what to transcribe and how to encode the text, passing through the processes of regularization and alignment, to every minute setting on how we express our data, are critical acts and require editorial judgement. Every single one of those interpretive acts places us closer to the possibility of presenting critically edited texts. *The CantApp: General Prologue* (Chaucer 2020; see also Bordalejo et al. 2020) contains our first edited text, a reader's edition. While working on it, I kept a separate record of the readings I would use in the production of a critical edition because I knew the time for including them as part of the Canterbury Tales Project work was near. Some might consider the inclusion of a critical text in our editions unnecessary because these editions present the differences between

documents. However, this would only be right if our understanding of a critical text was "traditional."

According to G. Thomas Tanselle, a critical edition is produced using editorial judgement and, more often than not, presents readings from various sources (Tanselle 1992, 27). My view of what a critical text represents is inspired by Klaus Wachtel and expressed by Robinson:

> ...the uses of the single text of the Nestle-Aland editions, and of the Münster *Editio Critica Maior*, do not depend on our accepting it as a precise reconstruction of a presumed first century original. There is another way of thinking of this text, which might reflect more closely the historical uncertainties about its origins and also provide a more fruitful perspective for his readers. One could think of this text a *the text that best explains all the extant documents*. The value of this text does not arise from its place as the endpoint of the editor's work, as the achieved and definitive reconstruction of the text as it may have existed at the moment of its composition. It has a different, less ambitious, but arguably more real value: it should be seen as the best starting point of the reader's own explorations of the text. (Robinson 2000)

Thus, my critical text will be one that explains the textual tradition as it stands, shows my understanding of variation, and serves as a gateway to the history of the work. Mine might seem the same as the theorization behind the Nestle-Aland text. Nestle-Aland also intends to explain all extant documents, but it is labelled an *initial* text. Instead, Vázquez and I describe critical editions as follows:

> The success of a critical edition relies on its ability to connect a system of data. With computer-assisted collation methods and full-text transcriptions, the process that leads to a critical text becomes comprehensive, thorough, and more transparent to the reader. In consequence, the critical text turns into a window through which we can observe the circumstances and the intervention of many of the agents that made it possible for us to engage with the texts. (Bordalejo and Vázquez 2020)

To put it even more clearly, when I talk about a critical text, I am not referring to the reconstruction of a lost archetype, but to the *construction* of a new, well-informed text that can help readers understand the relationships between extant witnesses; a text that functions as a gateway to the others. This text becomes in itself part of the textual tradition, not at the beginning of it (the top of an oriented stemma), but at the end: the latest version of the *Canterbury Tales*, created by the editor, who is no more than a knowledgeable, well-informed, studious scribe, who has used her understanding of textual relationships and her critical judgement to compose this new version as a tool to give herself and others a starting point for further study.

Because the Canterbury Tales Project has not yet offered a critical text as part of its editions, a hierarchy of variants has not been presented in the apparatus. Despite this, by privileging regularization towards the spellings in Hengwrt, readings have been normalized and other distinct systems pushed towards a vision of the text that is not quite real. In the wild, the *Canterbury Tales* variants come in all sorts of colours and flavours, not in the tame forms presented in the regularized apparatus, but in the fauvist diversity of their original spellings and bewildering word rearrangements. My critical text aims to highlight variance, not to mask it.

Bibliography

Alexanderson, Bengt. 2018. "Why Phylogenetic Methods Do Not Work Very Well in Textual Transmission." *Revue d'Histoire Des Textes* 13 (January): 383–410. https://doi.org/10.1484/J.RHT.5.114895.

Barbrook, Adrian C., Christopher J. Howe, Norman Blake, and Peter Robinson. 1998. "The Phylogeny of The Canterbury Tales." *Nature* 394 (6696): 839–839. https://doi.org/10.1038/29667.

Boc, Alix, Alpha B. Diallo, and Vladimir Makarenkov. 2012. "T-REX: A Web Server for Inferring, Validating and Visualizing Phylogenetic Trees and Networks" Nucleic Acids Research (W1): W573–W579.

Boc, Alix, Hervé Philippe, and Vladimir Makarenkov. 2010. "Inferring and Validating Horizontal Gene Transfer Events Using Bipartition Dissimilarity." *Systematic Biology* 59 (2): 195–211. https://doi.org/10.1093/sysbio/syp103.

Bordalejo, Barbara. 2002. "The Manuscript Source of Caxton's Second Edition of The Canterbury Tales and Its Place in the Textual Tradition of the Tales." De Montfort University. https://www.academia.edu/2987324/THE_MANUSCRIPT_SOURCE_OF_CAXTONS_SECOND_EDITION_OF_THE_CANTERBURY_TALES_AND_ITS_PLACE_IN_THE_TEXTUAL_TRADITION_OF_THE_TALES.

Bordalejo, Barbara. 2003. "The Phylogeny of the Order in the 'Canterbury Tales.'" PhD Thesis, New York, N.Y: New York University, Graduate School of Arts and Science. http://search.proquest.com/mlaib/docview/54046613/36CD5C07D51F43FAPQ/2.

Bordalejo, Barbara. 2016. "The Genealogy of Texts: Manuscript Traditions and Textual Traditions." *Digital Scholarship in the Humanities* 31 (3): 563–577. https://doi.org/10.1093/llc/fqv038.

Bordalejo, Barbara, and Peter M. W. Robinson. 2020a. "Wife of Bath's Prologue Regularized Apparatus." Zenodo. https://doi.org/10.5281/zenodo.3928655.

Bordalejo, Barbara, and Peter M. W. Robinson. 2020b. "Wife of Bath's Prologue NEXUS File." Zenodo. https://doi.org/10.5281/zenodo.3929763.

Bordalejo, Barbara, and Peter M. W. Robinson. 2019. "Manuscripts with Few Shared Significant Variants" 15 (July): 37–65. https://doi.org/10.5281/zenodo.3344880.

Barbara Bordalejo, Lina Gibbins, Richard North, & Peter Robinson. (2020). "Making an Edition in an App" (Version Pre-print). http://doi.org/10.5281/zenodo.3929929

Bordalejo, Barbara, and Adam Vázquez. 2020. “You’re Collating Just Fine and Other Lies You’ve Been Telling Yourself.” *Unpublished*, July. https://doi.org/10.5281/zenodo.3930286.

Cartlidge, Neil. 2001. “The Canterbury Tales and Cladistics.” Neuphilologische Mitteilungen 102:135–50.

Chaucer, Geoffrey. 1996. *Chaucer: The Wife of Bath’s Prologue on CD-ROM*. Edited by Peter Robinson. Cambridge: Cambridge University Press.

Chaucer, Geoffrey. 2004. *The Miller’s Tale on CD-ROM*. Edited by Peter Robinson. Canterbury Tales Project. Leicester: Scholarly Digital Editions. http://www.sd-editions.com/AnaAdditional/miller/images/millerhome.html.

Chaucer, Geoffrey. 2006. *The Nun’s Priest’s Tale on CD-ROM*. Edited by Paul Thomas. Leicester: Scholarly Digital Editions. http://www.sd-editions.com/NP/index.html.

Chaucer, Geoffrey. 2020. *The Tales of the Reeve and the Cook.* Edited by Thomas Farrell. Saskatoon: Inkless Editions. http://www.inklesseditions.com/TCP/Subscription/RE/?wit=Hg&page=50r&view=Image/Text.

Eagleton, Catherine, and Matthew Spencer. 2006. “Copying and Conflation in Geoffrey Chaucer’s Treatise on the Astrolabe: A Stemmatic Analysis Using Phylogenetic Software.” *Studies in History and Philosophy of Science Part A* 37 (2): 237–68. https://doi.org/10.1016/j.shpsa.2005.08.020.

Hanna, Ralph. 1996. *Pursuing History: Middle English Manuscripts and Their Texts*. Stanford Calif.: Stanford University Press.

Higgs, Paul G., and Teresa K. Attwood. 2005. *Bioinformatics and Molecular Evolution*. Hoboken, UK: John Wiley & Sons, Incorporated. http://ebookcentral.proquest.com/lib/usask/detail.action?docID=428071.

Howe, Christopher J., Ruth Connolly, and Heather F. Windram. 2012. “Responding to Criticisms of Phylogenetic Methods in Stemmatology.” *SEL Studies in English Literature 1500-1900* 52 (1): 51–67. https://doi.org/10.1353/sel.2012.0008.

Huson, D. H., and D. Bryant. 2006. “Application of Phylogenetic Networks in Evolutionary Studies.” *Mol. Biol. Evol* 23 (2): 254–67.

Maddison, David R, David L Swofford, and Wayne P Maddison. 1997. “NEXUS: An Extensible File Format Forr Systematic Information.” *Systematic Biology* 46: 32.

Manly, John M., and Edith Rickert. 1940. *The Text of the Canterbury Tales Studied on the Basis of All Known Manuscripts, Manly, John M.; Rickert, Edith ; with the Aid of Mabel Dean e. a.; with a Chapter on Illuminations by Margaret Rickert.* Chicago (Ill.): University of Chicago press.

Mendoza Straffon, Larissa, ed. 2016. *Cultural Phylogenetics*. Vol. 4. Interdisciplinary Evolution Research. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-25928-4.

Mooney, Linne R., Christopher J Howe, Adrian C Barbrook, and Peter M. W. Robinson. 2004. “Parallels between Stemmatology and Phylogenetics.” In *Studies in Stemmatology II*, edited by Pieter van Reenen, August den Hollander, and Margot van Mulken. Philadelphia, The Netherlands: John Benjamins Publishing Company. http://ebookcentral.proquest.com/lib/usask/detail.action?docID=623163.

Nei, Masatoshi, and Sudhir Kumar. 2000. *Molecular Evolution and Phylogenetics*. Oxford ; New York: Oxford University Press.

Page, Roderic D. M., and Edward C. Holmes. 1998. *Molecular Evolution: A Phylogenetic Approach*. Oxford ; Malden, MA: Blackwell Science.

Phillips-Rodriguez, Wendy J. 2007. "A Discussion about Textual Eugenics: Still Searching for the Perfect Mahābhārata?" In *Variants 6: Textual Scholarship and the Material Book*, 163–75. Brill | Rodopi. https://doi.org/10.1163/9789042028180_010.

Phillips-Rodriguez, Wendy J. 2010. "Some Considerations About Reading Stemmata." *Ecdotica* 9: 16–23.

Phillips-Rodriguez, Wendy J., Christopher J. Howe, and Heather F. Windram. 2009. "Chi-Squares and the Phenomenon of 'Change of Exemplar' in the Dyūtaparvan." In *Sanskrit Computational Linguistics*, edited by Gérard Huet, Amba Kulkarni, and Peter Scharf, 5402:380–90. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-00155-0_20.

Robins, William. 2007. "Editing and Evolution." *Literature Compass* 4 (1). https://doi.org/10.1111/j.1741-4113.2006.00391.x.

Robinson, Peter M. W. 1997. "A Stemmatic Analysis of the Fifteenth-Century Witnesses to The Wife of Bath's Prologue." In *The Canterbury Tales Projects Occasional Papers II*, edited by Peter M. W. Robinson and Norman Blake, 69–132. London: Office for Humanities Communication.

Robinson, Peter M. W. 2000. The One Text and the Many Texts. In *Literary and Linguistic Computing*, 15: 5-14.

Robinson, Peter M. W. 2003. "The History, Discoveries, and Aims of the Canterbury Tales Project." *The Chaucer Review* 38 (2): 126–39.

Robinson, Peter M. W. 2012. "The Textual Tradition of Dante's Commedia and the Barbi's Loci." *Ecdotica* 9: 1–32.

Robinson, Peter M. W. 2013. "The History of Scholarly Digital Editions, Plc." *Papers of The Bibliographical Society of Canada* 51 (1): 83-104. https://doi.org/10.33137/pbsc.v51i1.20763.

Robinson, Peter M. W. 2018. "Textual Communities XML File Format - Textual Communities - Wiki." June 27, 2018. https://wiki.usask.ca/display/TC/Textual+Communities+XML+file+format.

Robinson, Peter M. W. 2019a. "Creating and Implementing an Ontology of Documents and Texts (ADHO 2018) - Textual Communities - Wiki." January 31, 2019. https://wiki.usask.ca/pages/viewpage.action?pageId=1324745355.

Robinson, Peter M. W. 2019b. "Launch Document, 28 June 2018 - Textual Communities - Wiki." March 23, 2019. https://wiki.usask.ca/display/TC/Launch+document%2C+28+June+2018.

Robinson, Peter M. W. 2020. "Getting Materials from TC via URI/Linked Open Data - Textual Communities - Wiki." April 8, 2020. https://wiki.usask.ca/pages/viewpage.action?pageId=1306492976.

Robinson, Peter M. W., and O'Hara, Robert J. 1993. Computer-assisted Methods of Stemmatic Analysis (Version st). In *The Canterbury Tales Project Occasional Papers I*. edited by Peter M. W. Robinson and Norman Blake, 53–74. London: Office for Humanities Communication.

Roos, T., and T. Heikkila. 2009. "Evaluating Methods for Computer-Assisted Stemmatology Using Artificial Benchmark Data Sets." *Literary and Linguistic Computing* 24 (4): 417–33. https://doi.org/10.1093/llc/fqp002.

Roos, Teemu, and Yuan Zou. 2011. “Analysis of Textual Variation by Latent Tree Structures.” In *2011 IEEE 11th International Conference on Data Mining*, 567–76. Vancouver, BC, Canada: IEEE. https://doi.org/10.1109/ICDM.2011.24.

Smith, Catherine. 2019. *Itsee-Birmingham/Collation_editor_core 1.0.4*. Zenodo. https://doi.org/10.5281/zenodo.3539578.

Spencer, Matthew, Barbara Bordalejo, Peter Robinson, and Christopher J. Howe. 2003. “How Reliable Is a Stemma? An Analysis of Chaucer’s Miller’s Tale.” *Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing* 18 (4): 407–22.

Spencer, Matthew, Barbara Bordalejo, Li-San Wang, Adrian C. Barbrook, Linne R. Mooney, Peter Robinson, Tandy Warnow, and Christopher J. Howe. 2003. “Analyzing the Order of Items in Manuscripts of The Canterbury Tales.” *Computers and the Humanities* 37 (1): 97–109.

Spencer, Matthew, Barbara Bordalejo, Adrian C. Barbrook, Christopher How, Linne Mooney, & Peter Robinson. (2020, September 9). 'Gene order' analysis reveals the history of The Canterbury Tales manuscripts. Zenodo. http://doi.org/10.5281/zenodo.4021419

Spencer, Matthew, Elizabeth A Davidson, Adrian C Barbrook, and Christopher J Howe. 2004. “Phylogenetics of Artificial Manuscripts.” *Journal of Theoretical Biology* 227 (4): 503–11. https://doi.org/10.1016/j.jtbi.2003.11.022.

Swofford, D. L. 2003. *PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods)* (version 4.0a (build 167)). Sunderlan, Massachusetts: Sinauer Associates.

Tanselle, G. Thomas. 1992. *A Rationale of Textual Criticism*. University of Pennsylvania Press.

Wang, Li-San. 2001. “Exact-IEBP: A New Technique for Estimating Evolutionary Distances between Whole Genomes.” In *Algorithms in Bioinformatics*, edited by Olivier Gascuel and Bernard M. E. Moret, 2149:175–88. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-44696-6_14.

Windram, H. F., P. Shaw, P. Robinson, and C. J. Howe. 2008. “Dante’s Monarchia as a Test Case for the Use of Phylogenetic Methods in Stemmatic Analysis.” *Literary and Linguistic Computing* 23 (4): 443–63. https://doi.org/10.1093/llc/fqn023.

Windram, Heather F., Christopher J. Howe, and Matthew Spencer. 2005. “The Identification of Exemplar Change in the Wife of Bath’s Prologue Using the Maximum Chi-Squared Method.” *Literary and Linguistic Computing* 20 (2): 189–204. https://doi.org/10.1093/llc/fqi001.